Linguistic and Mixed Excitation Improvements on a HMM-based speech synthesis for Castilian Spanish

Xavier Gonzalvo, Joan Claudi Socoró, Ignasi Iriondo, Carlos Monzo, Elisa Martínez

GPMM - Grup de Recerca en Processament Multimodal
Enginyeria i Arquitectura La Salle. Universitat Ramon Llull
Quatre Camins 2, 08022 Barcelona (Spain).
{gonzalvo,jclaudi,iriondo,cmonzo,elisa}@salle.url.edu

Abstract

Hidden Markov Models based text-to-speech (HMM-TTS) synthesis is one of the techniques for generating speech from trained statistical models where spectrum and prosody of basic speech units are modelled altogether. This paper presents the advances in our Spanish HMM-TTS and a perceptual test is conducted to compare it with an extended PSOLA-based concatenative (E-PSOLA) system. The improvements have been performed on phonetic information and contextual factors according to the Castilian Spanish language and speech generation using a mixed excitation (ME) technique. The results show the preference of the new HMM-TTS system in front of the previous system and a better MOS in comparison with a real E-PSOLA in terms of acceptability, intelligibility and stability.

1. Introduction

One of the main problems of concatenative text-to-speech (TTS) systems is the degradation of quality when the database does not comprise the best units to be synthesized. Hence, larger databases are required for these kinds of systems. As the database grows up, it is more suitable to contain a unit closer to the target and more likely to have a better join [1]. In order to reduce errors, this database could become difficult to process. Therefore, a common solution is to use a limited domain context where text to be synthesized is under control (e.g. Virtual Weather man [2]).

Thence it follows that the final objective is to improve quality and naturalness in applications for general purpose. The main feature of the HMM-TTS is the statistical modelling of units producing a smoothed and natural speech that have been shown to be a possible advantage in front of the quality discontinuities in the concatenative systems [3]. Moreover, the main benefit of HMM-TTS is the capability of modelling voices in order to synthesize different speaker features, styles and emotions and perform interesting adaptations of speech [4]. Furthermore, HMM for speech synthesis could be used in new systems able to unify both approaches and to take advantage of their properties [5]. At this point, interesting work was presented by [6] to develop a fused system and last contributions have been presented in [7].

The aim of this paper is to present the advances throughout the development of a high-quality HMM-TTS for Castilian Spanish based on HTS engine [8]. Previous work for Spanish [9] identified the common problems that affect the HMM-TTS systems and other languages as well: vocoder, modelling accuracy and over-smoothing [7]. The following improvements are related to linguistic and vocoder issues which try to solve or alleviate these problems.

Firstly, the following linguistic features have been updated. In the one hand, the unit clustering has been upgraded using new contextual factors with respect to the previous approach [9], where the HMM training was presented to use a decision tree-based context clustering in order to improve models training. Also, clustering is able to characterize phoneme units introducing a counterpart approach with respect to English [3]. On the other hand, grapheme-to-phoneme conversion now uses a rule-based system to fix pronunciation errors instead of the Festival Spanish voice [10]. Secondly, synthesis quality has been increased by applying a mixed excitation (ME) technique using well defined models of the parametrized residual excitation [17]. The system is based on a source-filter model approach to generate speech directly from HMM itself. One of the drawbacks of these systems is the non ideal speech reconstruction due to the parametric representation of speech that the ME technique can solve by adding extra excitation parameters to the model.

This paper is organized as follows: Section 2 describes HMM system workflow and parameter training for spectrum, pitch, ME and duration. Section 3 concerns to synthesis process description. Section 4 presents measures, section 5 discusses results and final section presents the concluding remarks and future work.

2. HMM-based TTS system training

As in any HMM-TTS system, two stages are distinguished: training and synthesis. Figure 1 depicts the classical system training workflow (dotted lines stand for parameters modelled within the HMM).

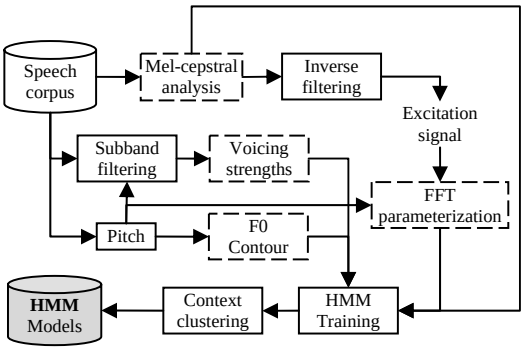

Figure 1: *Training workflow*

First, mel-cepstral analysis of the speech is performed. The first step estimates the HMM for isolated phonemes (each HMM represents a contextual phoneme) and each of these models will be used as an initialization of the contextual phonemes. Then, similar phonemes are clustered by means of a decision tree using contextual information and previously designed questions. Unseen units during the training stage can be synthesized using these decision trees. Each contextual phoneme HMM definition includes spectrum, state durations, F0, ME FFT parameters and the voicing strengths (VS) coefficients. During the analysis of these information, pitch is used for subband filtering and FFT parametrization.

Topology used is a 5 states left-to-right with no-skips. Each state is represented with 4 independent streams, one for spectrum, one for pitch and two more for mixed excitation part which comprises both FFT and VS. Each parameter is completed with its delta and delta-delta coefficients. The modelling information is structured in table 1.

Table 1: *Information modelled in the HMM.*

Feature vector streams			
c	Δc	Δ^2c	Spectrum
p	Δp	Δ^2p	F0
me	Δme	Δ^2me	FFT parameters for ME
v	Δv	Δ^2v	Voicing strengths

2.1. Spectrum modelling

The system is based on a source-filter model and spectrum parameters are modelled as multivariate Gaussian distributions [11]. Depending on the type and number of coefficients used on the vocoder, the quality of the synthetic speech can significantly vary. In this work, spectrum is updated to be modelled from 12^{th} to 24^{th} order mel-cepstral coefficients which generate speech with the MLSA (Mel Log Spectrum Approximation) filter [12]. The advantage of mel-cepstral in front of standard MFCC is that spectrum is better represented, so it gives a better performance of speech during synthesis [12]. Mel-cepstral has presented good results improving the basic HMM system in languages such as Arabic [13].

Last advances in high quality HMM-TTS used the STRAIGHT-based vocoding [14]. This analysis/synthesis technique is considered a high-quality solution initially used for speech morphing though it has been successfully applied to HMM-TTS (e.g. Blizzard 2005 [15]). Although it presents the advantage of performing pitch-adaptative spectral analysis, it was shown in [15] that MLSA filter was the most computationally efficient synthesis approach.

2.2. Mixed excitation

The aim of using a mixed excitation is to mimic the characteristics of natural human speech. It was first used in the LP vocoder (MELP) [16], a low bit rate speech coding and later integrated in a HMM-TTS for Japanese [17]. The reason for the vocoded speech quality is attributed mainly to the insufficiency of the binary source signal model which switches exclusively either the impulse train or the white noise. To solve this, the mixed excitation is implemented using a multi-band mixing structure.

As in the case of spectrum, STRAIGHT has also been used for the design of the mixed excitation as it weights a sum of a pulse train with phase manipulation and Gaussian noise. Other interesting schemes proposed the design of ME using wavelet [18].

The main information used to train the HMM is the following:

- Bandpass voicing strengths. The speech signal is filtered into five frequency bands considering a sample rate of 16k Hz [17] (see figure 1). The voicing strength in each band is estimated using normalized correlation coefficients around the pitch lag. In spite of correcting pitch estimation simultaneously with correlation, first the pitch is marked up and later, the correlation in each band is computed.
- Fourier magnitudes. In this work, the FFT parameters are the first thirty magnitudes of the centred pitch period of a 20ms excitation frame. The residual excitation is obtained by inverting the exponential filter transfer function [12] and filtering.

2.3. Pitch, mixed excitation and duration modelling

Pitch marks are crucial in order to obtain a good synthesis as they affect the representation of various parameters and the posterior training of the models. On the one hand, F0 contour is simultaneously modelled within the HMMs, hence estimated contour is dependent on the correctness of the pitch marks. On the other hand, mixed excitation FFT coefficients are estimated based on the determined pitch sequence. Thus, the Spanish corpus pitch analysis has been performed using an approach that automatically reduces the mark-up errors by using dynamic programming [19]. Moreover, this algorithm reduces discontinuities in the generated F0 curve for synthesis.

F0 model (table 1) is a multi-space probability distribution [11] that must be used in order to store continuous logarithmic values of the F0 curve and a discrete indicator for voiced/unvoiced. As in the case of spectrum, FFT magnitudes and voicing strengths are modelled as multivariate Gaussian distributions.

State durations of each HMM are modelled by a multivariate Gaussian distribution [20]. Its dimensionality is equal to the number of states in the corresponding HMM.

2.4. Phonetic data

The Spanish female voice was created from a corpus developed in conjunction with LAICOM [21]. Speech was recorded by a professional speaker in neutral emotion. Time boundaries segmentation was performed using an embedded HMM training, segmented and finally revised by speech processing researchers.

Phonetic labelling was performed in the previous work [9] using the Festival [10] Spanish voice. In order to resolve some incorrect transcriptions, a tested rule based approach (SinLib [22]) has been applied for text analysis in this work.

The grapheme-to-phoneme conversion has been extended from 31 to 36 units (see table 2) with one model of silence (types of silences are POS-tagged). It is important to notice that the system has the feature of a continuous transcription, so rules are applied between words (e.g. **/b**arko/ and /mi**B**arko/, translated as, “ship” and “my ship”).

- **Vowels**. Models for vowels are different either if they are stressed (capital letters) as also used in other approaches [23]. The system distinguishes various types of vowels:

semi-vowel, half open, open, closed and half closed including the main group of table 2.

- **Consonants**. New consonants (emphasized in bold) are used to avoid some pronunciation errors and improve intelligibility. Apart from the main groups, the system is also able to consider dental, velar, bilabial, alveolar, palatal, labio-dental, inter-dental, pre-palatal and voiced/unvoiced.

Table 2: *Castilian Spanish consonants and vowels inventory (SAMPA [24]).*

Vowels	
Frontal vowels	j,i,I,e,E,a,A
Back vowels	o,O,u,U,w
Consonants	
Plosive	p,b,t,d,k,g
Nasal	m,n,J,**N**,**M**
Fricative	**B**,f,tS,T,**D**,s,x,**G**
Lateral	l,L
Rhotic	R,r

2.5. Contextual factors

Input text is converted into a complete list of contextualized phonemes and each one is represented by a HMM. As the contextual information increases, HMMs will have less training data. To solve this problem during the training stage, similar units are clustered using a decision tree [11].

Extracted contextual information is language dependent and it serves as the features (attribute-value pairs) to construct the clustering decision trees. These trees are constructed using a set of questions designed in base of the contextual factors and the unit features using a yes/no based decision. Information referring to spectrum, F0, duration and ME is independently clustered by different trees.

Basically, the new approach in this work is focused on intonational improvement. English HMM-TTS included the ToBi tags which have been widely studied and applied to many systems [25]. In our case, we apply two groups of phonemes (Accentual group (AG) and Intonational group (IG)) in order to better represent the expressiveness. These parameters presented good results in a F0 estimator based on a machine learning approach applied to Spanish [9]. New information related to prosody events is the following:

- AG. Incorporates syllable influence and is related to speech rhythm. The type of AG is specified in Spanish as *agudo*, *plano*, *esdrújula* and *sobre-esdrújula* depending on the position of the accented syllable in the word.
- IG. Structure at this level is reached concatenating AGs. There are three types: interrogative, declarative and exclamative.
- AGs and IGs start/end flags.
- Syllable and word start/end flags.

New features are related to flags for syllable, words and intonational groups boundaries (SinLib system also controls these boundaries) and Part-of-speech (POS) that has been upgraded using Freeling [26] (a morphological engine). The following parameters are used to design the questions for the tree-based clustering and are presented in hierarchical order:

1. **Phonemes**. Current phone, left and before left phones and identical for the right side. Each kind of phoneme is labelled independently depending on the characteristics of table 2.
2. **AG**. The number of phonemes in current, previous and next AG; start/end flag and type of AG.
3. **IG**. Start/end flag and types of IG.
4. **Syllable**. Stress of current, previous and next syllables; position forward and backward of current syllable in current word and in current phrase; number of stressed syllables with respect to contextual syllables (this comprises 4 factors); vowel of the syllable and start/end flag.
5. **Word**. POS of the current, next and previous words; the number of syllables of current, next and previous words and position (forward and backward) of word in phrase and start/End flag.
6. **Phrase**. Number of syllables and number of words in current, previous and next phrases; positions (forward and backward) of current phrase in the utterance.
7. **Utterance**. Number of syllables, words and phrases in the utterance.

3. HMM-based TTS system synthesis

Figure 2 shows the synthesis workflow. Once the system has been trained, it has a set of phonemes represented by contextual factors. The first step is devoted to produce a complete contextualized list of phonemes from a text to be synthesized. Chosen units are converted into a sequence of HMM.

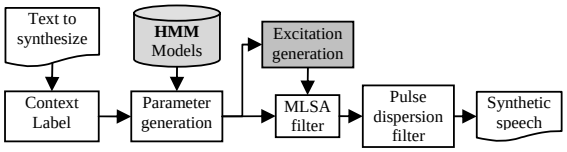

Figure 2: *Synthesis workflow*

Necessary parameters to synthesize are generated from the HMM using the algorithm proposed in [27]. The HMM is composed of the data and its Δ and Δ^2 features (see table 1). By taking into account the constraints between static an dynamic features, the algorithm avoid generating identical parameters for each state of the same HMM which results on an improved and smoothed speech envelope. Generated data are mel-cepstral, F0 and ME parameters. Duration is also estimated to maximize the probability of state durations.

Excitation signal is generated from the F0 curve, voiced and unvoiced information and the FFT parameters. Figure 3 presents the scheme to generate the mixed excitation (dotted lines indicates parameters generated from HMM). The pulse excitation is calculated from Fourier magnitudes using an inverse DFT of one pitch period in length. The bandpass filter for voiced and unvoiced parts are given by the sum of all the bandpass filter coefficients for the voiced and unvoiced frequency bands respectively. Voicing strengths are used to decide whether each filter coefficients belong to the voiced or unvoiced

part. The excitation is generated as the sum of the filtered periodic and noise excitations.

In order to reconstruct speech, the system uses spectrum parameters as the MLSA filter coefficients and excitation as the signal to filter. Finally, the obtained speech is filtered by a pulse dispersion filter which is a 130^{th} order FIR derived from a spectrally flattened triangle pulse based on a typical female pitch period. The pulse dispersion filter can reduce some of the harsh quality of the synthesized speech [16].

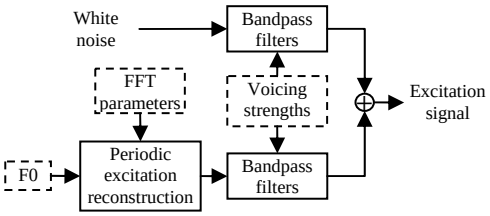

Figure 3: *Mixed excitation generation during the synthesis stage.*

One of the main problems during parameter generation is over-smoothing [28] that decreases the expressiveness and naturalness. Although the first solution would be to increase the size of the trees, its effect does not represent a substantial improvement in quality [9]. Another solution to improve the expressiveness could be to use an external F0 estimator though it can reproduce a forced intonation in some cases [9].

Last advances presented in [7] focus their study on reducing the error of the generated parameters. The HMM likelihood for a parameter trajectory generated by the conventional algorithm is too large compared with that for a natural one. This implies that is not only necessary to maximize the HMM likelihood [28]. For this case, minimum generation error (MGE) [29] or global variance (GV) [28] presented good results. GV introduces new constraints to the method of training and generation in order to avoid over-smoothing. The results reported were very good though at the moment is only showed to perceptually improve speech quality when applied to both mel-cepstral and F0.

4. Experiments

Experiments are conducted on a female corpus and evaluated using perceptual tests. The system was trained with HTS [8] using 620 phrases of a total of 833 (25% of the corpus is used for testing purposes). Contextual factors represent around 20000 units to be trained and around 5000 are unseen units.

Firstly, texts were labelled using contextual factors described in section 2.5. Then, HMMs are trained, decision trees for spectrum, F0, state durations and ME are built. Finally, HMM models are clustered. These trees are different among them because spectrum, F0 and states duration are affected by different contextual factors. Table 3 presents only two features to show the type of information in each tree. While spectrum tree is focused on phoneme features, excitation tree presents more high level information related to phrases (e.g. AG has increased the representation with respect to the spectrum tree).

It has been observed and discussed that RMSE is not a valid objective measure for F0 as it does not reflect real improvements showed by perceptual tests. For example, the generation algorithm considering GV usually causes larger errors compared with the conventional one [28] though GV increases the naturalness of synthesized speech. Meanwhile, subjective speech quality evaluation is generally seen to be the best measure of the aesthetic aspects [30] which is used to validate most of the TTS systems. Taking this into account, what follows presents a set of perceptual tests [1] to measure the improvements of the current HMM-TTS system.

Table 3: *Main contextual factors used for each tree.*

Feature vector	Contextual factors
Spectrum	Ph. 87%, AG 2%, Syll. 4%
Excitation	Ph. 45%, AG 16%, Syll. 10%
Durations	Ph. 76%, AG 8%, Syll. 5%
FFT	Ph. 21%, AG 11%, Syll. 28%
RV	Ph. 8%, AG 7%, Syll. 34%

In the first test, the systems with standard excitation (OLD-HMM) and the new system (ME-HMM) are evaluated. Figure 4 presents the preference of the new system in front of the old one. The effect of the ME (i.e. speech reconstruction buzzy is significantly reduced) is more important than the linguistic improvements. The preference tests evaluated single sentences by 15 listeners.

Figure 4: *Preference test for OLD-HMM and ME-HMM systems: (up) ME and linguistic improvement, (down) only linguistic improvements*

Once the new system has been validated, the second test (see figure 5) goal is to compare HMM-TTS systems with E-PSOLA [31] in terms of acceptability, intelligibility and naturalness. The perceptual comparisons were conducted using the same number of training sentences for both HMM-TTS and the E-PSOLA systems. Notice that the HMM-TTS systems model the F0 contour of a female voice with high variability (μ_{F0}=167 Hz, σ_{F0}=41 Hz) and the E-PSOLA version has real prosody from corpus as input.

The test was performed using a five steps (1-5) Mean Opinion Score (MOS) corresponding to the following quality evaluation: bad, poor, fair, good and excellent. The number of listeners were 25, most of them students of a technical degree and twenty phrases were randomly chosen for each system.

Different studies refer to acceptability as a measure of different components [30]. It is clear that in subjective user evaluations, at least intelligibility and naturalness play an important role. Subjective acceptability is not necessarily a simple consequence of intelligibility, and a distinction needs to be made between the aesthetic and functional aspects of synthetic speech.

1. **Acceptability**. Figure 5 shows that acceptability is higher for ME-HMM than for the other two systems, reaching a MOS of 2.8.
2. **Naturalness**. This measurement deals with quality and intonation as a measure of the extent to which a synthesizer sounds like a human [30]. In the one hand, the main

[1] See http://www.salle.url.edu/∼gonzalvo/hmm, for some synthesis examples

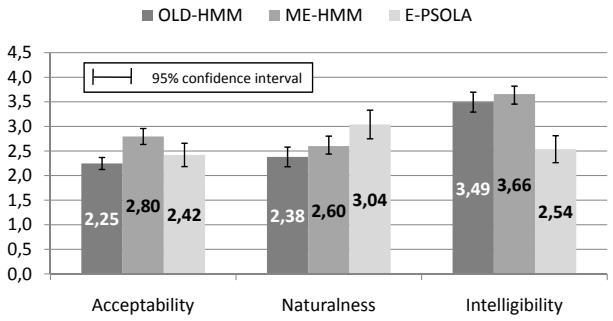

Figure 5: *Acceptability, intelligibility and naturalness MOS tests for ME-HMM, OLD-HMM and E-PSOLA systems.*

problem of the HMM-TTS is that produces a flat synthesis in some phrases. Moreover, although using a ME approach, the best example of a concatenative system still produces a better synthesis than the best HMM-TTS reconstruction [7]. On the other hand, E-PSOLA synthesis sounds more like a human but naturalness is affected by quality discontinuities. In any case, ME-HMM improves quality in comparison to the OLD-HMM due to the use of ME and new contextual factors (see section 2.5).

3. **Intelligibility**. This measurement marks the quality to distinguish the maximum number of words in a phrase. While E-PSOLA produces strong discontinuities that affect the comprehension of the phrases, HMM-TTS systems solve it by means of a smoother synthesis. This test also measures the effect of the linguistic changes (see section 2.4) with respect to the OLD-HMM.

Finally, as concluded for other languages (e.g. English [3] or European Portuguese [32]) HMM-TTS presents the most stable quality and although is less natural than E-PSOLA, it avoids quality discontinuities. In order to measure this, figure 6 shows the stability of the acceptability test in a bar graph. Notice that the E-PSOLA system is able to present more high-quality sentences but the probability of producing a bad synthesis is also higher than for the ME-HMM system. Stability of the ME-HMM system is then guaranteed thanks to a high probability "fair" zone.

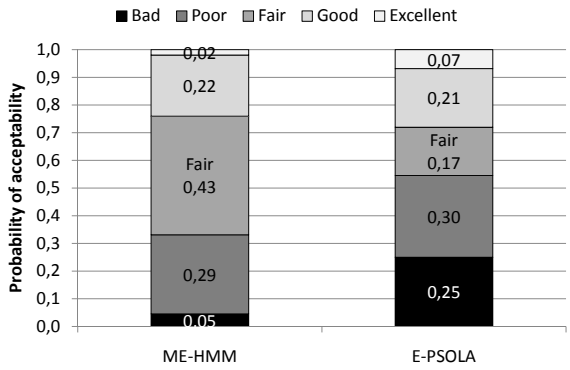

Figure 6: *Stability comparison based on the acceptability MOS results.*

5. Discussion

In order to analyse the concrete effect of the HMM-TTS, this work has presented a perceptual test in order to separate the factors that make a HMM-TTS preferable for general purposes applications controlling the length of the corpus.

In the one hand, the advantage of the HMM-TTS systems is its ability to maintain the synthesis quality for any text to be synthesized and the main drawback is the naturalness of the final produced speech. Using a HMM-TTS provides a high intelligibility system, that could even be more independent of corpus label errors than a standard concatenative system. In fact, the perceptual results could justify one of the possible aspects to make the acceptability be higher for system based on HMM-TTS, that is, the intelligibility and a quality able to reduce the vocoded speech.

Therefore, HMM-TTS systems used in a non limited domain applications provide stability. The intelligibility test could be the main reason because results have shown that smooth speech with a high intelligibility is preferable though a concatenative system still provided a higher naturalness.

6. Conclusions and future work

This work has presented the improvements on a Spanish HMM-TTS based on HTS updating new phonetic information, appending the AG and IG to contextual factor and integrating a ME scheme. With a set of tests we have compared the performance against a concatenative synthesis system. Subjective measures presented the advance of the system in terms of acceptability, intelligibility, naturalness and stability. The results have shown that the HMM-TTS for Spanish presents a better intelligibility and the ME reduced the buzzy vocoder quality. Also acceptability and stability of the system has presented an advantage in front of other kinds of synthesis in general purposes application.

HMM-TTS produces a flat synthesis caused by a smooth F0 contour and mel-ceptral parameters estimation. The conclusion from the results is that the HMM-TTS system is more suitable due to produce a continuous and more stable synthesis. However, although naturalness has been improved with regards to the previous system, it is still a lack and more expressiveness is still desirable. In this aspect, it seems to be necessary to integrate a parameter generation using minimum error to gain expressiveness and naturalness. New techniques and vocoders (e.g. Harmonic-Noise Model or STRAIGHT) have presented successful results in TTS systems, so a logical step would be to compare its performance with our current system. Moreover, it would be interesting to shape the HMM generated F0 contour with an external F0 estimation using an extended version of the system presented in the last approach [9].

Voice transformation and conversion techniques will be applied in the future. Finally, perceptual tests have been used to measure the subjective quality of the system. Due to RMSE is not a correct measure to objectively measure the improvements of the systems, it would be desirable to propose a new objective measure to evaluate the HMM-TTS systems quality that could also be extended to other types of synthesis. Voice quality descriptors could deal with this topic in the future.

7. Acknowledgements

This work has been developed under SALERO (IST FP6-2004-027122). This document does not represent the opinion of the European Community, and the European Community is not re-

sponsible for any use that might be made of its content.

8. References

[1] Black, A., "Perfect synthesis for all of the people all of the time", Proc. of IEEE SSW, 2002.

[2] Alías, F., Iriondo, I., Formiga, Ll., Gonzalvo, X., Monzo, C. and Sevillano, X., "High quality Spanish restricted-domain TTS oriented to a weather forecast application", Proc. of Interspeech, 2005.

[3] Tokuda, K., Zen, H. and Black, A.W., "An HMM-based speech synthesis system applied to English", Proc. of IEEE SSW, 2002.

[4] Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T., "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR", Proc. of ICASSP, pp.805–808, May 2001.

[5] Donovan, Robert E. and Woodland, P. C., "A hidden Markov-model-based trainable speech synthesizer", Computer Speech and Language, vol.13, pp.223–241, 1999.

[6] Taylor, P., "Unifying Unit Selection and Hidden Markov Model Speech Synthesis", Proc. of Interspeech, 2006.

[7] Black, A., Zen, H. and Tokuda, K., "Statistical Parametric Speech Synthesis", Proc. of ICASSP, pp.1229–1232, 2007.

[8] Tokuda, K., Zen, H., Yamagishi, J., Masuko, T., Sako, S., Black, A.W and and Nose, T., "The HMM-based speech synthesis sysmte (HTS)", http://hts.ics.nitech.ac.jp

[9] Gonzalvo, X., Iriondo, I., Socoró, J.C., Alías, F. and Monzo, C., "HMM-based Spanish speech synthesis using CBR as F0 estimator", Proc. of NoLISP, 2007.

[10] Black, A. W., Taylor, P. and Caley, R., "The Festival Speech Synthesis System", http://www.festvox.org/festival

[11] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi T. and Kitamura, T., "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis", Proc. of Eurospeech, pp. 2374–2350, 1999.

[12] Fukada, T., Tokuda, K., Kobayashi, T. and Imai, S., "An adaptive algorithm for mel-cepstral analysis of speech", Proc. of ICASSP, pp.137140, 1992.

[13] Ossama, A-H., Sherif Mahdy, A. and Mohsen, R., "Improving Arabic HMM based speech synthesis quality", Proc. of Interspeech, pp.1332-1335, 2006.

[14] Kawahara, H., Estill, Jo. and Fujimura, O., "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT", MAVEBA, 2001.

[15] Zen, H. and Tomoki, T., "An overview of nitech HMM-based speech synthesis system for blizzard challenge 2005", Proc. of Interspeech, pp.93–96, 2005.

[16] McCree, A. V. and Barnwell III, T. P. , "A mixed excitation LPC vocoder model for low bit rate speech coding", IEEE Trans. Speech and Audio Processing, vol.3, no.4, pp.242-250, Jul. 1995.

[17] Yoshimura, T., Tokuda, K., Masukom,T., Kobayashi, T. and Kitamura, T., "Mixed excitation for HMM-based speech Synthesis", Proc. of Eurospeech, pp.2259-2262, Sept. 2001.

[18] Aoki, N., Ifukube, T. and Takaya, K., "Implementation of MELP vocoder using lifting wavelet transform", Proc. IEEE Region 10 Conf. TENCON, pp.194-197, Sept. 1999.

[19] Alías, F., Monzo, C. and Socoró, J.C., "A Pitch Marks Filtering Algorithm based on Restricted Dynamic Programming", Proc. of Interspeech, 2006.

[20] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., "Duration Modeling in HMM-based Speech Synthesis System", Proc. of ICSLP, vol.2, pp.29–32, 1998.

[21] Iriondo, I., Socoró. J.C., Formiga, L., Gonzalvo X., Alías F. and Miralles P., "Modeling and estimating of prosody through CBR", Proc. of JTH, 2006. (In Spanish)

[22] http://www.salle.url.edu/tsenyal/english /recerca/areaparla/tsenyal_software.html

[23] Lambert, T. and Breen, A., "A database design for a TTS synthesis system using lexical diphones", Proc. of Interspeech, pp.1381–1384, 2004.

[24] Llisterri, J. and Mario, J.B., "Spanish adaptation of SAMPA and automatic phonetic transcription", UPC, ESPRIT PROJECT 6819 (SAM-A Speech Technology Assessment in Multilingual Applications), 1993

[25] Black, A. and Hunt, A., "Generating F0 contours from ToBI labels using linear regression", Proc. of ICSLP, vol 3, pp.1385–1388, 1996.

[26] Atserias, J., B. Casas, E. Comelles, M. Gonzlez, L. Padr and M. Padr, "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library", Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genoa, Italy. 2006.

[27] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", Proc. of ICASSP, pp.1315-1318, 2000.

[28] Toda, T. and Tokuda, K., "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis", IEICE Transactions, Vol. E90-D, No. 5, pp.816–824, 2007.

[29] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis", Proc. of ICASSP, pp.89–92, 2006

[30] Lampert, A., "Evaluation of the MU-TALK Speech Synthesis System", ICT Report, 2004

[31] Iriondo, I., Alías, F., Sanchis, J., Melenchón, J., "A Hybrid Method Oriented to Concatenative Text-to-Speech Synthesis", Proc. of Eurospeech, vol. 4, pp.2953–2958, 2003.

[32] Barros, M. J., Maia, R., Tokuda, K., Freitas, D. and Resende Jr., F. G., "HMM-based European Portuguese Speech Synthesis", Proc. of Interspeech, pp.2581-2584, 2005.